Gene Flow–Dependent Genomic Divergence between *Anopheles gambiae* M and S Forms

David Weetman,*,[1] Craig S. Wilding,[1] Keith Steen,[1] João Pinto,[2] and Martin J. Donnelly[1]

[1]Vector Group, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

[2]UEI Parasitologia Médica/Centro de Malária e outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa. Lisbon, Portugal

***Corresponding author:** E-mail: dweetman@liv.ac.uk.

Associate editor: Matthew Hahn

Abstract

Anopheles gambiae sensu stricto exists as two often-sympatric races termed the M and S molecular forms, characterized by fixed differences at an X-linked marker. Extreme divergence between M and S forms at pericentromeric "genomic islands" suggested that selection on variants therein could be driving interform divergence in the presence of ongoing gene flow, but recent work has detected much more widespread genomic differentiation. Whether such genomic islands are important in reproductive isolation or represent ancestral differentiation preserved by low recombination is currently unclear. A critical test of these competing hypotheses could be provided by comparing genomic divergence when rates of recent introgression vary. We genotyped 871 single nucleotide polymorphisms (SNPs) in *A. gambiae* sensu stricto from locations of M and S sympatry and allopatry, encompassing the full range of observed hybridization rates (0–25%). M and S forms were readily partitioned based on genomewide SNP variation in spite of evidence for ongoing introgression that qualitatively reflects hybridization rates. Yet both the level and the heterogeneity of genomic divergence varied markedly in line with levels of introgression. A few genomic regions of differentiation between M and S were common to each sampling location, the most pronounced being two centromere–proximal speciation islands identified previously but with at least one additional region outside of areas expected to exhibit reduced recombination. Our results demonstrate that extreme divergence at genomic islands does not simply represent segregating ancestral polymorphism in regions of low recombination and can be resilient to substantial gene flow. This highlights the potential for islands comprising a relatively small fraction of the genome to play an important role in early-stage speciation when reproductive isolation is limited.

Key words: *Anopheles gambiae*, malaria vector, speciation island, single nucleotide polymorphism, hybridization.

Research article

Introduction

How genomes evolve during speciation is a fundamental but poorly understood question in evolutionary biology (Wu 2001; Wu and Ting 2004; Nosil et al. 2009a; Butlin 2010). Identifying the number of genes involved in reproductive isolation can give insight into early-stage speciation processes. If fewer loci drive divergence, differentiation between incipient species may accrue more quickly and with a greater level of gene flow because selection coefficients operating on each locus can be higher (Nosil et al. 2009b). This could reduce the requirement for allopatric separation and allow diversification to continue between sympatric taxa, consequently increasing potential for speciation. As reproductive isolation advances, identifying variants that were critical for initial divergence becomes difficult as high levels of differentiation may be observed at loci across the genome that are not involved (Via 2009). Consequently, systems at an early stage of divergence are valuable models. Sympatric races of insects, particularly those exhibiting a spectrum of reproductive isolation, represent some of the best current models for research into speciation genomics. Of these, the African malaria mosquito *Anopheles gambiae* sensu stricto appears to be one of the most promising (Turner and Hahn 2010) due to the availability of a near-fully assembled genome sequence and advanced genomic tools (Holt et al. 2002; Lawniczak et al. 2010; Neafsey et al. 2010).

The *A. gambiae* complex comprises at least seven morphologically indistinguishable species (Coluzzi et al. 2002) with varying levels of range overlap and reproductive isolation (Ayala and Coluzzi 2005). Within *A. gambiae* sensu stricto, detection of fixed differences in intergenic and internal transcribed ribosomal DNA spacers (IGS and ITS rDNA) near the centromere of the X chromosome led to the definition of "molecular forms" termed M and S (della Torre et al. 2001; Gentile et al. 2001), which occur in sympatry throughout much of west and central Africa (della Torre et al. 2005). Though the molecular forms were originally thought to covary with chromosomal forms, defined by sets of inversion polymorphisms on chromosome 2 (Coluzzi et al. 2002), it is now clear that all common inversion polymorphisms are shared between the molecular forms and probably predate the division of M and S (Costantini et al. 2009; Simard et al. 2009; White et al. 2009). Interform hybrids are identified using single-locus pericentromeric X-linked polymerase chain reaction (PCR) diagnostics, which target either a substitution

in the IGS-rDNA (Fanello et al. 2002) or a form-specific *Sine200* transposable element insertion (Santolamazza et al. 2008a). Information on the survival of hybrids in the wild is lacking, but F_1 hybrids are fully fertile in the laboratory (Diabate et al. 2007). Typical estimates of hybridization rates between forms range between 0% and 1.5% (Yawson et al. 2004; della Torre et al. 2005; Costantini et al. 2009; Simard et al. 2009) though much higher rates (up to 25%) have been documented recently at the western edge of *A. gambiae*'s range (Caputo et al. 2008, 2011; Ndiath et al. 2008; Oliveira et al. 2008). Complementing these data, behavioral evidence suggests incomplete assortative mating (Tripet et al. 2001; Diabate et al. 2009), microsatellites often show low interform differentiation (Lanzaro et al. 1998; Taylor et al. 2001; Lehman et al. 2003; Yawson et al. 2007), and insecticide resistance mutations have introgressed between forms (Weill et al. 2000; Djogbénou et al. 2008; Weetman et al. 2010). Evidence for a genomic mechanism that could reconcile fixed differences on the X chromosome in the apparent presence of substantial gene flow was provided by a landmark genomewide microarray genotyping study (Turner et al. 2005). Large concentrations of single feature polymorphisms differentiated the M and S forms near the centromeres of chromosomes X and 2L, but differentiation appeared relatively low elsewhere.

In summary, these data suggested that the molecular forms of *A. gambiae* sensu stricto are at an early stage of speciation driven by genes in the highly differentiated genomic islands, with the rest of the genome homogenized by ongoing gene flow (Turner et al. 2005; Turner and Hahn 2007). Indeed, ongoing or recent gene flow is critical for this "speciation island" model because otherwise the regions of extreme divergence could be ancestral polymorphisms, which segregate between molecular forms and are preserved by low recombination (Noor and Bennett 2009; Turner and Hahn 2010; White et al. 2010). With any plausible estimate of effective population size, the hybridization rates in most sampled areas suggest interform gene flow in excess of that required to prevent extreme divergence in the absence of selection (Wright 1931; Slatkin 1987). However, selection against F_1 hybrids in the wild can be very strong (e.g., McBride and Singer 2010), and it is possible that a large number of the hybrids detected are F_1 hybrids rather than more advanced crosses or backcrosses. This could mean that gene flow between M and S forms is much lower than it might appear (White et al. 2010). By contrast, introgression of insecticide resistance mutations between molecular forms, followed by subsequent spread, provides compelling evidence for backcrossing of F_1 hybrids. Nevertheless, such events could be extremely rare, with subsequent increases in allele frequency driven by strong selection on these insecticide resistance-conferring mutations (Lynd et al. 2010), rather than repeated introgression. Recent findings of an additional pericentromeric island of divergence (White et al. 2010) and high differentiation throughout the genome (Lawniczak et al. 2010; Neafsey et al. 2010; Weetman et al. 2010) suggest that ongoing interform gene flow might indeed be very low in *A. gambiae* and question whether the major genomic islands are key or incidental to speciation (Turner and Hahn 2010; White et al. 2010).

Determining the mode of speciation in *A. gambiae* has important implications for public health in Africa. Behavioral and ecological differences between the M and S forms are likely to influence malaria epidemiology by spatial and temporal extension of transmission ranges (Lehman and Diabate 2008). Interform differences in insecticide resistance are also common (Santolamazza et al. 2008b; Weetman et al. 2010), though resistance is increasing in previously-susceptible M forms in some areas (Dabiré et al. 2009, Lynd et al. 2010). Therefore, whether speciation in *A. gambiae* is a dynamic process, which could occur over very short evolutionary timescales, is a question of both academic and public health relevance. Genomewide studies to date have focused on samples from the central and eastern parts of the M and S sympatric range where hybridization rates are low to moderate. Populations from the western edge of *A. gambiae*'s range, where hybridization rates are much higher, might, however, offer the best opportunity to distinguish speciation models.

In this study, we genotyped samples from across *A. gambiae*'s range to investigate the relationship between rates of recent introgression and the nature of genomewide differentiation. Initially, we examined whether the classification to molecular form according to one or two single-marker diagnostics has a wider genomic basis throughout the species range, even where interform hybridization rates are very high. We then tested sequential predictions, which would distinguish the genomic island model of adaptive divergence with gene flow between M and S, from a null hypothesis of ancestral segregation, with minimal recent gene flow. First we sought evidence for recent introgression to determine whether hybrids have enough reproductive potential for contemporary interform gene flow to compare qualitatively with hybridization rates. Second, if levels of introgression differ between samples, the proportion of the genome segregating between forms should decrease as introgression rates increase, yielding a more clustered profile of differentiation (Wu 2001; Wu and Ting 2004; Nosil et al. 2009b; Via 2009). Third, if X-marker–diagnosed M and S status has a wider genomic basis and loci that segregate between the two forms are fewer when introgression is higher, at least some genomic regions of extreme interform differentiation should be conserved across locations in a common mosaic of adaptive divergence (Via and West 2008; Via 2009).

Materials and Methods

Sample Sites and Identification of Species and Molecular Form

Full details of sampling, DNA extraction, and genotyping methodologies are given elsewhere (Müller et al. 2008; Weetman et al. 2010), but briefly, samples were collected as larvae from Yaoundé, Cameroon (03° 52′ N, 11° 31′ E) in July to August 2006 and Dodowa, Ghana (05° 53′ N, 00°

06′ W) in October to November 2006 and raised to adults. The sampling protocol scaled the number of larvae taken to habitat size to reduce the possibility of sampling siblings. Indoor-resting adult females were collected (by aspiration) from houses in Tororo, Uganda (00° 41′ N, 34° 10′ E,) in November 2008 and (by trapping in untreated bednets) from houses in Antula, Guinea-Bissau (11° 53′ N, 15° 34′ W) in August to September 1993. All samples were confirmed morphologically as *A. gambiae sensu lato* A PCR–restriction fragment length polymorphism (RFLP) molecular diagnostic of X chromosome rDNA variation (Fanello et al. 2002) was used to determine species identity within the *A. gambiae sensu lato* group and *A. gambiae* sensu stricto molecular form (M or S). A portion of samples were also screened using an alternative assay, which detects a form-specific insertion of a *Sine200* element on the X chromosome (Santolamazza et al. 2008a). Though unproblematic elsewhere, interpretation of the rDNA IGS PCR–RFLP in Guinea-Bissau was frequently ambiguous (see also Caputo et al. 2011), and all Guinean samples were diagnosed as M or S form using the *Sine200* diagnostic, which as a single rather than multicopy marker is more readily scored. Samples encompassed the range of known hybridization rates, which appear temporally stable over multiple collections points. In Cameroon, no hybrids have been detected in over 15,000 specimens genotyped (Simard et al. 2009), though introgression of the *kdr 1014F* mutation from S to M forms demonstrates some recent gene flow (Etang et al. 2009; Weetman et al. 2010); in Ghana, the rate of hybridization is 0.25–0.5% (Yawson et al. 2004, 2007; Weetman et al. 2010; Egyir-Yawson A, Weetman D, and Donnelly MJ, unpublished data); in Guinea-Bissau, hybridization rates are 19–25% Oliveira et al. 2008; Caputo et al. 2011).

Single Nucleotide Polymorphism Genotyping

Following whole genome amplification, DNA samples were screened on an Illumina Beadstation GX using the Illumina GoldenGate assay according to the manufacturer's protocols. Genotyping was performed with a 1,536 single nucleotide polymorphism (SNP) array, originally designed primarily to cover approximately 260 candidate genes potentially related to insecticide resistance. Nevertheless, the array also genotyped over 300 control SNPs and provided coverage throughout much of the genome. Moreover, very few of the genes have been found to be insecticide-resistance associated or to show evidence of selection via patterns of linkage disequilibrium (Weetman et al. 2010). SNPs originated from sequencing of mixed populations of M and S forms (Wilding et al. 2009) and from those identified during whole genome sequencing of M and S form colonies from Mali (Lawniczak et al. 2010). All SNPs chosen to populate the array were polymorphic in both M and S forms. Genotyping arrays were scored using Beadstudio v3.2 (Illumina Inc.): Of the 1,536 SNPs on the array, 871 could be scored reliably (i.e., showed good clustering of genotypes, >80% call rate and no evidence for null alleles) in all sample collections and were polymorphic in at least one. Only these SNPs were used in the present analysis (genomic locations of the 871 SNPs scored are shown in supplementary figure s1, Supplementary Material online). Only samples typing as pure forms using the above PCR diagnostics (i.e., not hybrids) were genotyped. Since we detected no hybrids in our collection from Cameroon and only two from Ghana, this had negligible impact on the sample but did result in exclusion of approximately 20% of the sample from Guinea-Bissau. Final sample sizes for analysis were Cameroon M form ($N = 673$), Cameroon S form ($N = 62$), Ghana M form ($N = 29$), Ghana S form ($N = 769$), Uganda S form ($N = 214$), Guinea-Bissau M form ($N = 40$), and Guinea-Bissau S form ($N = 40$).

Data Analysis

Individual-based clustering of multilocus genotypes was performed using the Bayesian algorithm implemented by BAPS 5.3 (Corander and Marttinen 2006, 2008). Each run was repeated many times to check that the optimal clustering solution had been obtained. Analyses were repeated using different data sets involving subtraction of SNPs from different chromosomes: 1) X, 2 and 3; 2) 2 and 3; 3) 3, rather than for each chromosome separately owing to the lower number of SNPs, and resultant low statistical power, on the physically short chromosome X. BAPS 5.3 was also used to identify individuals from locations of M and S sympatry whose genomes showed evidence of significant mixture of M and S. This algorithm first estimates which multilocus genotypes show evidence of mixture and the proportion of the genome attributed to each source population, followed by simulation of multilocus genotypes from allele frequencies to determine the posterior probability that putatively mixed genotypes could be found in the resident population (Corander et al. 2006, 2008). Sufficient simulations (2,000–16,000 depending on sample size) were performed to permit identification of individual genotypes as significantly mixed following correction for multiple testing based on the false discovery rate criterion (Benjamini and Hochberg 1995). This is expected to be conservative with respect to type I error rate for the BAPS mixture analysis (Corander et al. 2006). We ran a two-step analysis to distinguish recent introgression from ancestral retention/historic gene flow. We first determined the mixture proportions, and associated probabilities, for sympatric M and S samples and determined which were significantly mixed. We then repeated the analysis but including allopatric samples (minimum distance 3,000 km), which could not have contributed recent immigrants, to determine whether the sympatric source of admixture remained the most likely (adjudged as >50% source of the immigrant proportion). Only those individuals meeting the criterion of significant admixture from a majority sympatric source were deemed to exhibit evidence of recent introgression; for others, the null hypothesis of ancestral mixture was accepted. As with the clustering analysis, we first applied analyses using all SNPs and then repeated excluding data from chromosomes X and 2 since regions of reduced recombination might either repel or retain extended

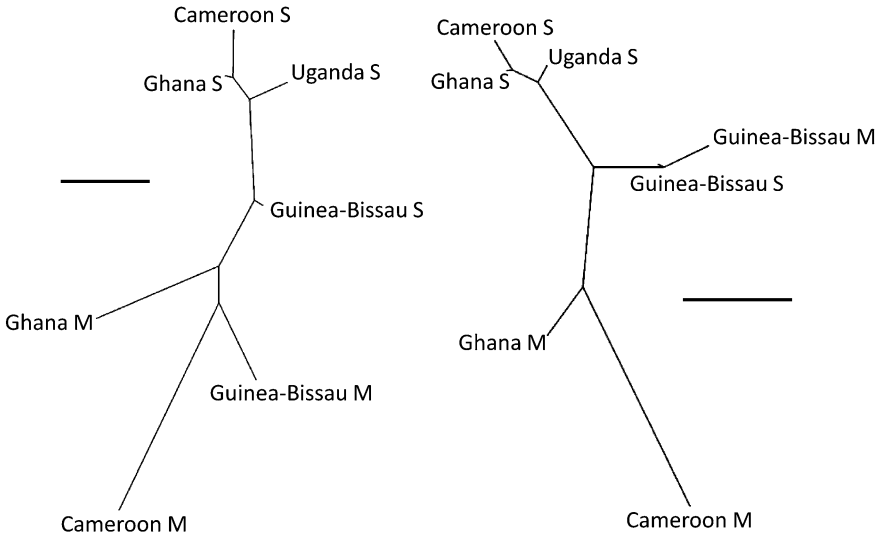

FIG. 1. Unrooted neighbor joining trees showing differentiation among M and S forms from different locations. Trees are based on linearized F_{ST} calculated using (*A*) SNPs from all chromosomes ($N = 871$) and (*B*) SNPs from only chromosome 3 ($N = 301$). Thick horizontal lines to the side of each plot show a scale of $F_{ST}/(1 - F_{ST}) = 0.05$. Note the high similarity of trees in *A* and *B*, with the exception of the M-form sample from Guinea-Bissau.

portions of immigrant chromosomes and potentially cause bias in estimates.

Genepop v4 (Rousset 2008) was used to compute F_{ST} values. Differentiation at individual SNPs in pairwise comparisons between populations was computed using Fisher's exact test following rescaling of all samples to the minimum scored at any SNP ($N = 44$ chromosomes). This permitted direct comparisons among samples by correcting for the strong dependence of *P* values on sample size. Bonferroni correction for multiple testing was applied to determine the significance of probabilities (critical $\alpha = 0.05/N$ polymorphic SNPs in comparison). FDist (Beaumont and Nichols 1996) implemented by Lositan (Antao et al. 2008) was used to identify loci unlikely to have evolved neutrally under equilibrium expectations, using 10,000 coalescent simulations. Baseline F_{ST} for the simulations was calculated from all the SNP data following iterative exclusion of outliers at the 5% level by Lositan, with observed F_{ST} values considered significant if they exceeded the 0.95 confidence interval of simulations. The number of SNPs expected to cooccur as outliers in population pairs and in all three populations was calculated from the numbers observed in single populations. In cases where exact tests involved contingency tables exceeding 2×2 cells, the software RxC was used to calculate probabilities by permutation (Miller 1997). Haploview 4.1 (Barrett et al. 2005) was used for reconstruction of population haplotypes (within form and within sampling location) to compute linkage disequilibrium (measured as r^2). Neighbor-joining trees were produced from linearized F_{ST} values ($F_{ST}/(1 - F_{ST})$ by Phylip 3.68 (Felsenstein 2008) and drawn using FigTree 1.3.1 (Rambault 2006).

Results

Discrimination of M and S Form Genomes

Genetic structure among sample collections based on SNPs from all chromosomes revealed distinct division of the molecular forms with differentiation more pronounced among M forms (fig. 1*A*). We repeated the analysis using SNPs from chromosome 3 to avoid possible influence of extended regions of reduced recombination found on the chromosomes X and 2 (Stump et al. 2005; Pombi et al. 2006, 2008). Differentiation was generally reduced (supplementary table s1, Supplementary Material online), but there was virtually no effect on tree topology apart from the repositioning of Guinean M forms to the branch with Guinean S forms, intermediate in position between other M and S collections (fig. 1*B*). Individual-based Bayesian cluster analysis using SNPs from all chromosomes produced distinct partitioning of most sample collections, with the only major overlap being between the S forms from Cameroon and Ghana (fig. 2*A*). M and S forms from all locations were almost wholly distinct and their clusters separated, with only 0.2% of assignments to clusters of the alternate molecular form (fig. 2*A*; supplementary fig. s2*A*, Supplementary Material online). As classification of M and S forms is based on pericentromeric single-marker diagnostics on the X chromosome, linkage disequilibrium with multiple other X chromosome polymorphisms could create bias. Therefore, we repeated the analysis excluding all SNPs located on the X chromosome: This resulted in only a marginal increase in mixed clustering in Guinea-Bissau (two M forms were assigned to the Guinean S cluster) but not elsewhere. Repetition of the clustering analysis using only chromosome 3 SNPs had no impact

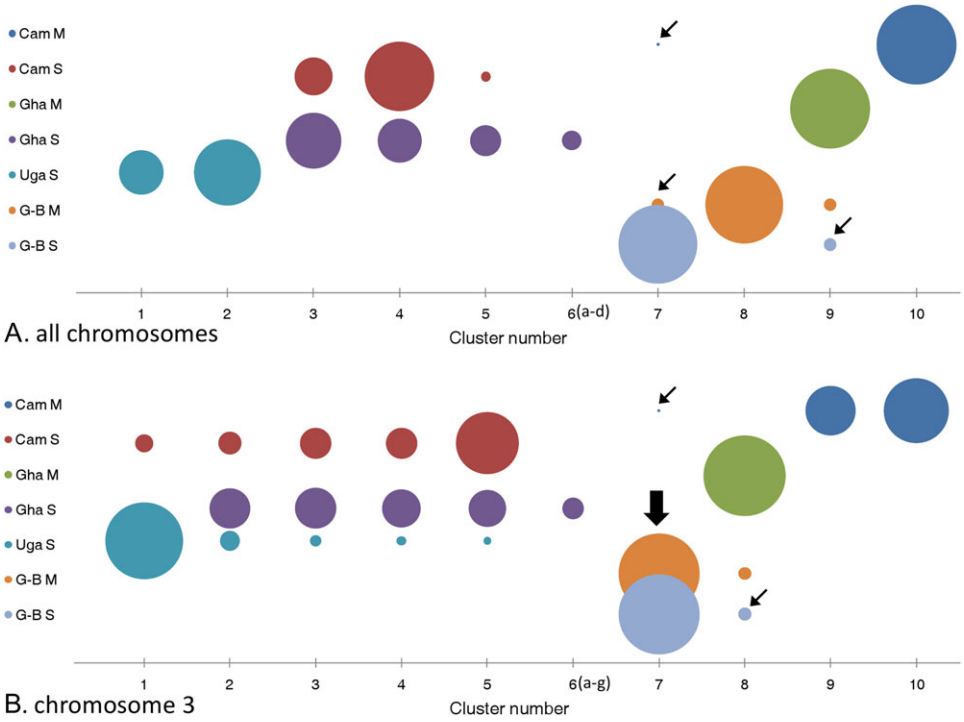

Fig. 2. Individual-based (BAPS) cluster analyses of individuals from all sample locations. Bubble sizes represent the proportion of individuals from a sample site present in a cluster. Cluster numbers are not intended to correspond between (*A*) and (*B*), though memberships may overlap considerably (e.g., cluster 9 in [*A*] is identical to cluster 8 in [*B*]). Arrows highlight the exceptions to the distinct clustering of M and S individuals. Thin arrows denote single individuals; the thick arrow indicates many individuals (as clusters 7 + 8 in [*A*] merge into cluster 8 in [B]). To simplify presentation, multiple very small clusters (all Ghanaian S forms), have been merged into cluster 6 in each plot.

on partitioning of the M and S forms from Uganda, Cameroon, or Ghana but resulted in a collapse of segregation between forms in Guinea-Bissau, which merged into a single cluster intermediate between other M and S clusters (fig. 2*B*; supplementary fig. s2*B*, Supplementary Material online). Thus, in general, M and S forms were strongly differentiated and readily partitioned, but for the Guinea-Bissau, collection discrimination of M forms from S was dependent on the inclusion of chromosomes with extended regions of reduced recombination.

Recent Introgression

The proportion of individuals exhibiting significantly mixed genomes and whose interform mixture was assigned to the sympatric source (see "Materials and Methods") differed depending on the chromosomal data set used. In Cameroon and Ghana, chromosome 3–based estimates were about 1/3 and 2/3, respectively, of those based on all SNPs (fig. 3*A*). The difference is probably attributable to introgression of the strongly selected insecticide resistance–associated *kdr 1014F* mutation from S to M forms in both collections. *Kdr 1014F* is located within the voltage-gated sodium channel gene near the centromere of chromosome 2L within an extended haplotype represented by many SNPs on our array (Lynd et al. 2010; Weetman et al. 2010). As this haplotype is otherwise absent from the M forms in our collections, *kdr* introgression can generate a strong signal of mixture, even if the overall immigrant

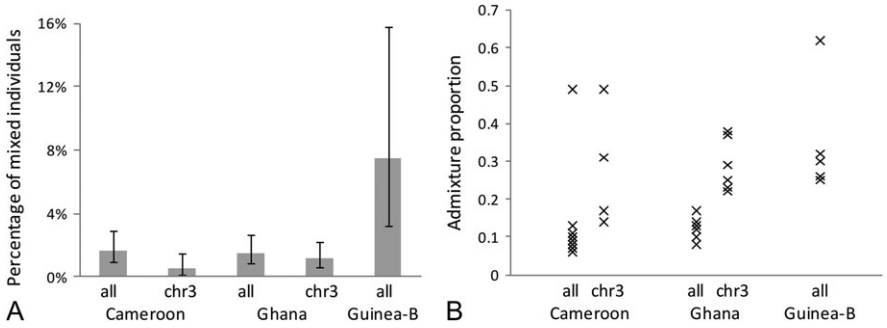

Fig. 3. Evidence of recent introgression between sympatric M and S forms. *A*. Percentage of individuals (+/− 95% binomial confidence intervals) showing significant evidence for genomic mixture attributable to the alternate molecular form in sympatry (see text). *B*. Admixture proportions for each significantly mixed individual. In *A* and *B*, data are shown for analyses based on all SNPs and chromosome 3 SNPs, with the exception of Guinea-Bissau for which very low chromosome 3 differentiation prevented analysis.

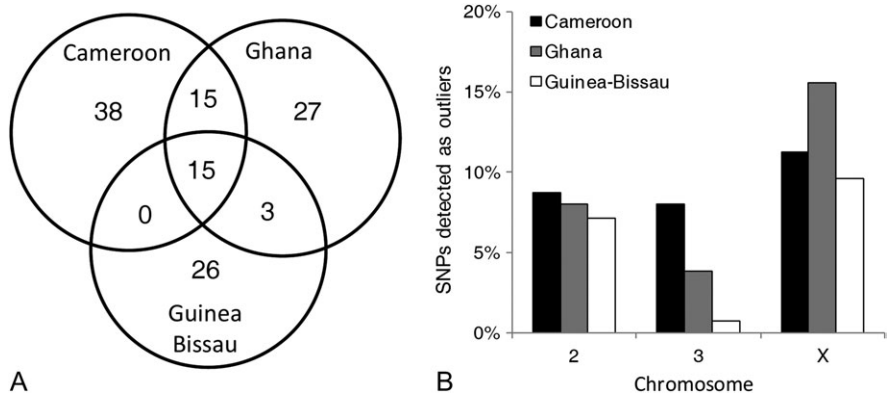

FIG. 4. F_{ST} outlier analyses of M versus S comparisons. (*a*) Number of SNPs detected as outliers in each location and replicated as outliers in multiple locations. (*b*) Percentage of SNPs classed as outliers on each chromosome.

proportion of the genome is relatively low. Consistent with this, far fewer M form individuals possessing (introgressed) *kdr* were detected as mixed from chromosome 3 SNPs (supplementary table s2, Supplementary Material online), and almost all admixture proportions were higher (fig. 3*B*). The *kdr 1014F* mutation was only detected in M forms in Cameroon and Ghana 3–4 years before our samples were collected (Yawson et al. 2004; Etang et al. 2006). Therefore, the mixture events detected from chromosome 3 SNPs must be of relatively recent origin because only one of the individuals possessing introgressed *kdr* was detected as mixed in the chromosome 3 analysis (supplementary table s2M, Supplementary Material online). Lack of interform differentiation in the Guinea-Bissau collection (see fig. 2*B*) precluded mixture analysis for chromosome 3 SNPs, but absence of the *kdr* mutation in this collection would presumably limit bias in mixture estimation evident in Cameroon and Ghana.

As expected from the higher rates of hybridization, the percentage of individuals exhibiting significantly mixed genomes was significantly higher in Guinea-Bissau than in Cameroon and Ghana (fig. 3*A*). The proportion of the genome originating from the alternate form for mixed individuals in Guinea-Bissau (all chromosome data) and Ghana (chromosome 3 data) were generally high (all >0.2; fig. 3*B*), consistent with very recent crossing events, such as backcrossing of hybrids to parental forms. The few mixed ancestry individuals in Cameroon (chromosome 3 data) showed higher variation in the proportion of the genome from the other molecular form, but at least one individual exhibited admixture consistent with very recent crossing (approximately 50:50 M:S; fig. 3*B*). There was no evidence for significant directional bias in interform introgression in any collection (exact test, minimum $P = 0.32$).

If we assume that counting only significantly mixed individuals for which a sympatric source explained the majority of the admixture proportion was effective in excluding ancestral mixture, the mixture data can provide a rough estimate of a single parameter, the contemporary effective migration rate (m). We calculated m from the sum of (significant and sympatric) genome proportions divided by total sample size (e.g., Hänfling and Weetman 2006). Estimates, based on chromosome 3 SNPs, are $m = 0.0015$ for Cameroon, $m = 0.0031$ for Ghana, and based on all SNPs, $m = 0.025$ for Guinea Bissau, noting the important caveat that the estimate of m for Guinea-Bissau could be downwardly biased as hybrids diagnosed using the *Sine 200* marker (see "Materials and Methods") were not included. Therefore, in support of our first prediction, estimates of interform migration rates are qualitatively similar to hybridization rates, which are about 100-fold lower in Ghana than in Guinea-Bissau and extremely low in Cameroon where hybrids have never been detected (references above).

Location of Genomic Divergence Between Forms

The number of SNPs adjudged excessively differentiated relative to neutral expectations (and thus consistent with divergent selection) in sympatric comparisons is shown in figure 4*A*. Distribution of outliers across chromosomes differed among populations ($P = 0.008$), with heterogeneity the strongest in Guinea-Bissau, where only a single outlier was located on chromosome 3 (fig. 4*B*). Physical clustering of outliers among populations is also evident in figure 5. The numbers of SNPs significantly differentiated between M and S forms varied among samples: Cameroon 16.6%; Ghana 12.6%; Guinea-Bissau 3.6% ($\chi^2_2 = 65.4$, $P << 0.001$; note that sample sizes were equalized for this analysis), supporting our second prediction that fewer loci should segregate between the two forms in sympatric samples where introgression was higher. Allopatric comparisons between M forms and Ugandan S forms, between which gene flow is implausible, yielded similar profiles of differentiation to sympatric M–S comparisons in Cameroon and Ghana (figure 5). Pericentromeric X chromosome loci showed a near-identical level of differentiation in sympatric and allopatric interform comparisons for M forms from Guinea-Bissau. However, elsewhere in the genome and most markedly on chromosome 2, Ugandan S and Guinean M forms were far more strongly differentiated than the sympatric Guinean M and S. This is consistent with earlier results that it is the S rather than M

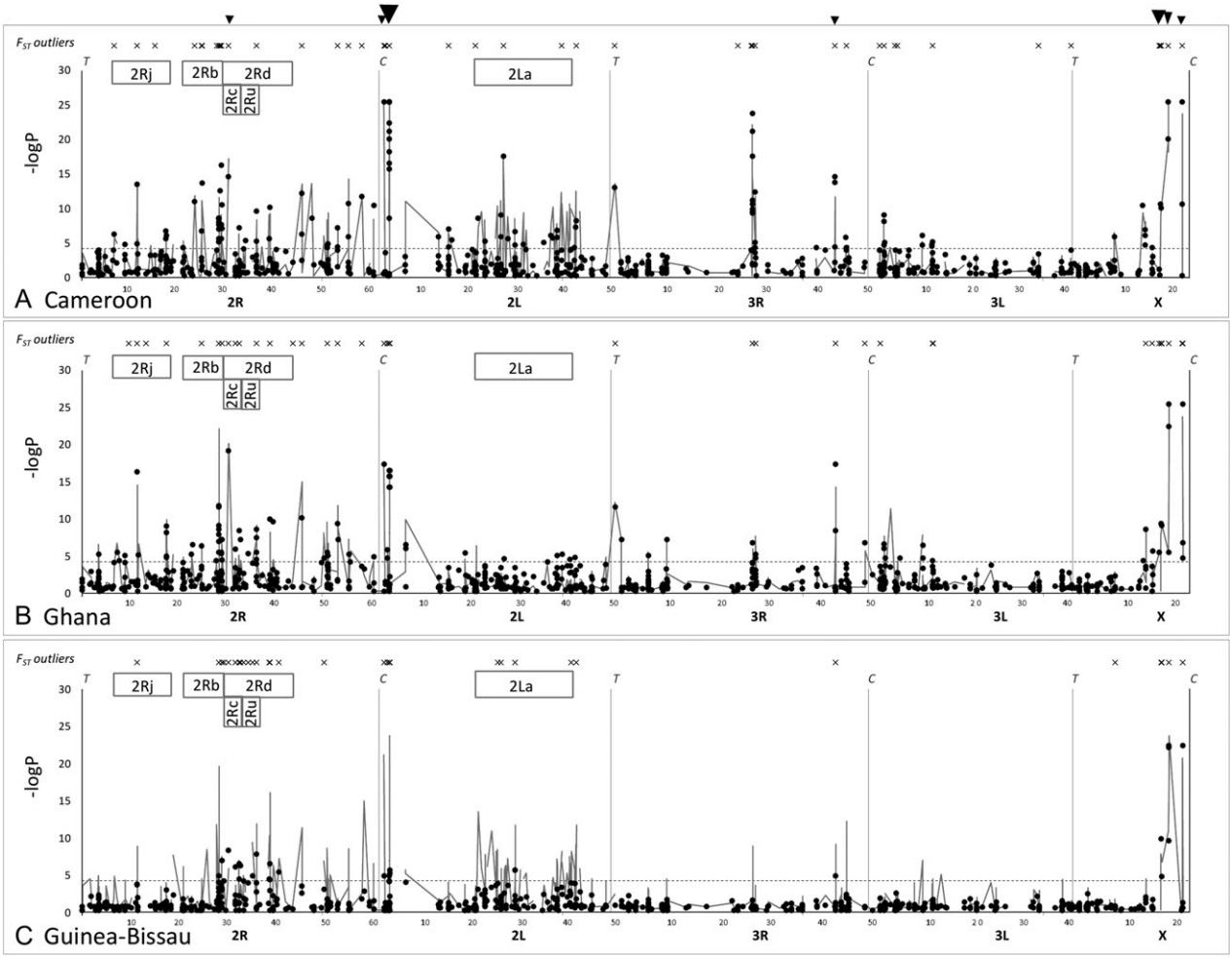

Fig. 5. Genomic differentiation of M and S forms. Probabilities from exact tests for differentiation at each SNP are plotted against a linear physical scale for equalized sample sizes from each sample pairing: Dots show sympatric M–S comparisons; lines show allopatric comparisons of M forms with S forms from Uganda. Horizontal line indicates the bonferroni-corrected α level. Vertical lines indicate centromeres (C) and telomeres (T). Boxes show inversion regions. F_{ST} outlier SNP positions are indicated by crosses at the top of each plot; outliers found in all populations are indicated by arrows at the top of the figure (size proportional to number of SNPs where clustered).

forms in Guinea-Bissau that exhibit an unusual pattern of differentiation compared with other samples of the same form (see fig. 1*A*, supplementary fig. s2*A*, Supplementary Material online).

To determine whether genotypic assignment to M or S form depended on the outlier SNPs, we repeated the clustering analysis, permitting two clusters for each sampling location. All individuals from Cameroon were assigned correctly to M and S form whether or not outliers were included. For Ghanaian samples, removal of outliers resulted in a 15% drop in correct assignment (from 100%). However, in Guinea-Bissau, M and S forms, for which there was 99% correct assignment with outlier SNPs included, could not be discriminated upon removal of outlier SNPs (i.e., only a single cluster was found), despite there being fewer outliers than in the other samples. Surprisingly, given the large number of outlier SNPs present within the 2Rc/d/u polymorphic inversion region (see fig. 5), especially in Guinea-Bissau, removal of outliers located within any inversion region did not alter assignment success in any sample collection.

Our third prediction was that genomic regions of interform divergence should be conserved across locations if the X-marker–diagnosed M and S status has a common origin and some wider genomic basis. Fifteen SNPs were scored as outliers in all sympatric comparisons (fig. 4*A*), far more than the number expected from numbers of singleton outliers (exact test, $P << 0.001$). These 15 consistent outlier SNPs were found on four of the five chromosome arms. The majority of these SNPs are within or very close to the 2L and X chromosome centromeric heterochromatin, within the locations of the previously identified islands of divergence between M and S forms or within a polymorphic inversion (supplementary figure s3, Supplementary Material online). In each of these areas, recombination is expected to be reduced. By contrast, the consistently replicated outlier SNP position on chromosome 3R at approximately 45.9 Mb, which has not been identified previously, is not within or close to heterochromatin (Sharakova et al. 2010) or any common polymorphic inversion (Pombi et al. 2008).

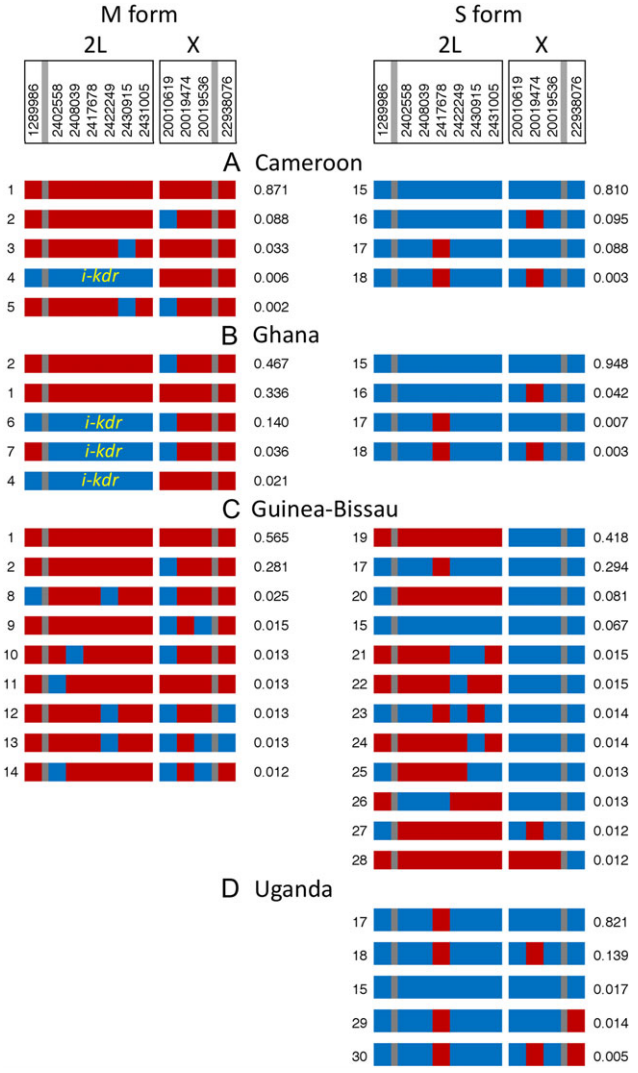

FIG. 6. Haplotypes for the 2L and X centromere–proximal outlier groups. SNPs are named by base position on each chromosome. Alleles more common in M forms are shown as red blocks and those more common in S forms as blue. Physical separation is indicated by a break in the haplotype. Haplotype number is to the left of each block and frequency to the right. Haplotypes predicted to occur in at least one chromosome in the sample are shown. *i-kdr* = introgressed *kdr 1014F*.

Evidence for Recombination Within and Between Major Islands of Divergence

We performed haplotype analysis to identify possible recombination events within and among the outlier blocks on chromosomes 2L and X, which are within the previously described genomic islands (fig. 6). Specifically, we aimed to identify how the occurrence of haplotypes that were primarily characteristic of one molecular form varied in frequency in the alternate form across sample collections. Since the pattern of interform differentiation on the X chromosome suggested a steady rise toward the outliers located at or beyond 20 Mb, we focused on these SNPs (supplementary figure s3, Supplementary Material online). For S forms, there was little evidence to suggest introgression in Cameroon and Ghana (fig. 6*A* and *B*), with all haplotypes entirely form-characteristic, and the only M-type alleles (colored red in fig. 6) also present in Ugandan haplotypes (fig. 6*D*). Interestingly, the most common haplotype in Uganda was rare in S forms in Cameroon and Ghana, perhaps reflecting selection for different *kdr* mutations since it differed only within the 2L section. In Cameroonian and Ghanaian M forms, it is plausible that an allele at the edge of the X island (SNP 20010619 in fig. 6) might have originated from S forms, but absence of an entirely allopatric M sample precludes any conclusion. Indeed, the only clear introgression was of haplotypes on which the *kdr 1014F* mutation was present, which resulted in mixed haplotypes (S-like 2L, with M-like X). More haplotypes were found in Guinean M forms, but it is notable that >80% were the same two that predominated in the M forms in Cameroon and Ghana, suggesting that introgression from S forms is relatively uncommon. By contrast, the most common haplotypes in Guinean S forms were actually a mixture of those characteristic of an M form 2L island haplotype block but an S-like X island block (fig. 6*C*).

Table 1 summarizes the evidence for interisland recombination. There is no evidence for recombination between the 2L and X islands in S forms from Cameroon and Ghana for which genotype frequency distributions do not differ significantly from those found in Uganda. With the exception of *kdr* introgression–related genotypes, there is negligible evidence for interisland recombination in M forms other than Guinea-Bissau, where a moderate frequency of X-M/M, 2L-M/S genotypes was found. In Guinean S forms, the 2L island appears entirely to dissociate freely from the X (table 1). In conclusion, we detected little evidence of interform recombination within or between the X and 2L genomic islands in Cameroonian or Ghanaian samples, beyond that attributable to *kdr*-related introgression. In Guinea-Bissau interform recombination appears to be more common within the 2L genomic island and between the 2L and X islands. Yet the marked bias from M to S forms in genomic island introgression contrasts with the lack of any such bias detected in the recent introgression (admixture) analysis (above).

Table 1. Genotypes at Form-Characteristic Markers for the 2L and X Genomic Islands

		2L-M/M	2L-M/S	2L-S/S	Diff
Cam	X-M/M	660	9*k*	0	a
Cam	X-S/S	0	0	59	b
Gha	X-M/M	20	7*k*	1*k*	c
Gha	X-S/S	0	1	765	b
GB	X-M/M	32	5	0	c
GB	X-S/S	8	22	9	d
Uga	X-S/S	2	0	214	b

NOTE.—Genotypes for the Sine X markers and 2L_≈1.3 Mb SNP (the most divergent 2L outlier) are shown, with frequencies in each population: Cam = Cameroon, Gha = Ghana, GB = Guinea-Bissau, and Uga = Uganda. M/M indicates alleles characteristic of M forms, S/S of S forms. Genotype distributions of populations sharing letters in the "diff" column are not significantly different. *k* = *kdr* introgression event.

Discussion

Discriminating Between Models of Divergence

The M and S molecular forms of *A. gambiae* sensu stricto have been the subject of intense recent research, leading to major advances in our understanding of their intrinsic genetic, behavioral, and ecological differences (e.g., Turner et al. 2005; Lehman and Diabate 2008; Costantini et al. 2009; Simard et al. 2009). Almost all work to date has focused on the eastern and central parts of their range of overlap, where M and S genomes can be distinguished readily (Slotman et al. 2006; Esnault et al. 2008). With samples spanning most of the range of *A. gambiae* sensu stricto, our data extend previous results by showing that the genomes of M and S forms are at least partially distinct, even when X-marker–diagnosed hybridization rates are extreme. This might be expected if F_1 hybrids, or perhaps hybrids in general, are very strongly selected against in the wild (e.g., McBride and Singer 2010). Our results suggest that this is not the case; at least some interform introgression appears to be occurring in all populations at levels qualitatively comparable with hybridization rates (i.e., Cameroon and Ghana $<<$ Guinea-Bissau). This was our first prediction and its support argues against the retention of ancestrally divergent genomic islands by neutral processes alone (Turner et al. 2005; Noor and Bennett 2009; Turner and Hahn 2010).

Our second prediction was that the proportion of the genome segregating between forms should decrease when introgression rates are higher (Wu 2001; Wu and Ting 2004; Nosil et al. 2009b; Via 2009). Not only was interform divergence higher in Cameroon than in Ghana and much greater than in Guinea-Bissau but also the distribution of F_{ST} outliers and their contribution to M and S differentiation differed markedly. In Cameroon, significant differentiation was present along all chromosomes. SNPs identified as F_{ST} outliers were also distributed across all chromosomes, and the M and S partitioned perfectly even if these SNPs were excluded from clustering analyses. Differentiation and outlier SNPs were somewhat less uniformly distributed in Ghana, and omission of outliers did reduce the capacity of the data to distinguish M and S by 15% of correct assignments. Each observation was much more extreme for the Guinean samples. Little of the genome was differentiated, including only one of the 286 polymorphic SNPs on chromosome 3 (compared with 22 in Ghana and 37 in Cameroon). F_{ST} outliers were significantly clustered, and despite comprising only 5% of SNPs, if they were omitted, it was no longer possible to accurately assign individuals to molecular form. Thus, prediction two was supported by the data, meeting the expectation of Wu's (2001) gene-centric model of speciation with gene flow. In addition, this is consistent with the expectation that genes critical for early-stage divergence might be more readily apparent when gene flow is higher (Via 2009).

As noted recently elsewhere (Lawniczak et al. 2010; Neafsey et al. 2010; Weetman et al. 2010), interform divergence across the genome was more widespread than suggested by data of Turner et al. (2005), which reflects primarily the different resolution of the techniques applied (Turner and Hahn 2010; White et al. 2010). Yet, the extent of M–S differentiation we found in Cameroon, where 0.3% of SNPs exhibited fixed differences, is incompatible with report of Lawniczak et al. (2010) of 8% fixed differences between Malian M and S forms, which would suggest the absence of regular gene flow. Hybrids are found regularly in Mali but not to date in Cameroon, so lower gene flow between the forms in the latter location is an implausible explanation for the discrepancy. However, it is important to note that in the Lawniczak et al. study, the individuals sequenced were drawn from laboratory colonies, which typically harbor only a fraction of the diversity of wild populations (Norris et al. 2001). This likely led to a major upward bias in differentiation relative to natural populations. Indeed, patterns of differentiation between wild-caught Malian M and S forms (Neafsey et al. 2010) are much more consistent with the low level of fixed differences in our data set.

Our final prediction was the most specific: that some genomic regions of interform divergence should be conserved across locations (Via and West 2008; Via 2009); this would argue for involvement of a core set of loci in divergence of the molecular forms. Support for this prediction comes from the observation that 15 outlier SNPs were common in M–S comparisons across populations, which far exceeded the number expected (15 observed and <1 expected). The most obvious areas of shared divergence were the X and 2L pericentromeric areas, the genomic islands identified previously (Turner et al. 2005). If it is accepted that SNPs on the X chromosome in positions <20 Mb are outliers because of divergence hitchhiking (Via and West 2008; Via 2009), as appears consistent with a pattern of gradual decrease in differentiation (see supplementary figure s3, Supplementary Material online), the size of both the 2L and the X islands correspond closely with those obtained using a genotyping method with different resolution (Turner et al. 2005; White et al. 2010).

The 2L and X islands are within areas of known or expected reduced recombination (e.g., Pombi et al. 2006) as was the novel outlier within the 2Rc/d/u inversion region. The final consistent outlier detected—at approximately 46 Mb on chromosome 3R—is not within or close to the centromeric heterochromatin or any common inversion region. Unfortunately, data on recombination rates in this region (Zheng et al. 1996; Pombi et al. 2006) are too limited at present to conclude with confidence that background recombination in this area is similar to the genomic average. Multiple other outliers were shared between Cameroon and Ghana (supplementary table s3, Supplementary Material online), including a SNP in the small island on chromosome 2R identified in Cameroon by Turner et al. (2005), the importance of which was subsequently downgraded because it was not detected in Mali (Turner and Hahn 2007). Of these outliers, a few were close to significance in Guinea-Bissau, including at least one in an area where there is no expectation of reduced

recombination (supplementary table s3, Supplementary Material online). We therefore contend that our third prediction is met and that there are shared areas of divergence between M and S evident across sampling locations, even when interform gene flow is at its highest. Moreover, as also demonstrated by the individual clustering analysis, these are not limited to the pericentromeric region of the X chromosome, where the form-diagnostic markers are located (della Torre et al. 2001; Santolamazza et al. 2008a).

Maintenance of Genomic Islands

Our data suggest that multiple islands are associated with and might be causal in advancing reproductive isolation, which is probably to be expected, given the somewhat multifarious nature of phenotypic differentiation between *A. gambiae* M and S forms. Assortative mating occurs via swarm segregation (Diabate et al. 2009) and sophisticated mate recognition within swarms (Pennetier et al. 2010), but there are additional phenotypes that differ between the molecular forms that may also enhance divergence. Although phenotypic differences such as variation in larval growth rate and predator avoidance (Lehman and Diabate 2008; Gimmoneau et al. 2010) have no direct connection with assortative mating, they may enhance differentiation by segregating the forms in time and space (see Costantini et al. 2009; Simard et al. 2009). In Guinea-Bissau, where divergence is lowest and gene flow highest, we detected only a few islands, and it is possible that limited divergence at specific regions, which segregate M and S strongly elsewhere (supplementary table s3, Supplementary Material online>), contributes to the reduced reproductive isolation. For example, a strongly divergent region extending approximately 1.7 Mb from the centromere of chromosome 3L has been observed in several low gene flow locations (Neafsey et al. 2010; White et al. 2010) but was not detected in our Guinea-Bissau data. Our array has low SNP coverage in this region of the genome, but a marker approximately 1.9 Mb from the centromere of chromosome 3L was identified as an outlier in both Cameroon and Ghana (fig. 5; supplementary table s3, Supplementary Material online). In addition, a PCR–RFLP diagnostic (White et al. 2010) for the 3L centromeric region does not segregate strongly among the molecular forms in Guinea-Bissau (Caputo et al. 2011; Weetman D, Wilding CS, Steen K, Pinto J, and Donnelly MJ, unpublished data). From these date, we infer that the 3L genomic island is either absent or weakly differentiated in Guinea-Bissau.

Support for the predictions we made strongly suggests that variation in divergence between M and S forms across the genome more readily fits a model of heterogeneous genomic divergence with gene flow (Wu 2001; Turner et al. 2005) than a model of retention of ancestral divergence with minimal gene flow (Turner and Hahn 2010; White et al. 2010). However, reduced background recombination might still play an important role (Noor and Bennett 2009; Nosil et al. 2009b). Turner et al. (2005) used coalescent simulations to investigate whether the fixed differences they observed at the major X and 2L speciation islands could be explained solely by reduced recombination. Based on a hypothesized recombination rate of 0.1 cM Mb^{-1} and migration of $4N_em = 10$ (estimated indirectly from F_{ST}), the authors concluded that divergent selection must also play an important role. The recombination rate applied is about half the value reported subsequently for the X chromosome in the region 20–23 Mb (Pombi et al. 2006), and the value of N_em also appears slightly conservative based on our more direct estimation from the Cameroon data set and an effective population size, N_e, of at least a few thousands (e.g., Lehman et al. 1998). Therefore if Turner et al. (2005) used conservative parameter estimates, as appears to be the case, there is even stronger support for a key role for selection in maintaining the X chromosome island.

There was scant evidence for recombination within the X chromosome island. In Guinea-Bissau (M forms) where our genomic estimates of recent introgression rates were substantial, rare alleles more characteristic of the alternate molecular form were detected but at frequencies little higher than found in Uganda, where there is no realistic chance of interform recombination (see fig. 6C and D). Recombination rate estimates are unavailable for pericentromeric regions of chromosome 2L, but haplotype composition suggested a quite different scenario. In Cameroon and Ghana, the form specificity of 2L island haplotypes was only disrupted in cases where the *kdr 1014F* insecticide resistance mutation had introgressed. As a result of the dominance of the introgressed 2L haplotype in the source S form populations (91% in Cameroon and 99% in Ghana), our data can provide little indication of the frequency of *kdr* introgression events. Thus, based on the results from Cameroon and Ghana, we cannot exclude the possibility of a single introgression event into a region of extremely low background recombination, with frequency subsequently inflated by insecticidal selection (Lynd et al. 2010). Fortunately, the haplotype patterns in Guinea-Bissau permit greater insight of the integrity of the 2L island. Multiple 2L haplotypes were detected, with M-characteristic haplotypes dominant in S forms, suggestive of rather free recombination. This implies that background recombination is unlikely to be low enough to explain the segregation of M and S at the 2L island seen in other locations (present study; White et al. 2010). Similarly, the strong covariation between form-specific marker polymorphisms for the 2L and X islands (present study; White et al. 2010) was completely absent from Guinea-Bissau S forms, ruling out any intrinsic physical restrictions on recombination across the centromeres of distinct chromosomes (White et al. 2010; see also Caputo et al. 2011). The final evidence that the 2L island is maintained by selection comes from Guinea-Bissau M forms. Though more 2L haplotypes were present, almost 85% were the same two most common in Cameroon and Ghana, despite the apparently dramatic inequality in effective migration rates and a lack of detectable directional bias in gene flow from M to S. Thus, gene flow had a far greater impact on the integrity of the 2L genomic island in the Guinean S forms than in the M; a pattern

consistent with much stronger selection for form-specific polymorphisms in the M than in the S forms.

Conclusions

Our results describe a dynamic gene flow–dependent process of genomic divergence in *A. gambiae*, with a key role for selection acting on several genomic regions. The largest of these appear be the 2L and X islands identified originally by Turner et al. (2005); though gene densities therein are so low (Holt et al. 2002), it is quite possible that far smaller regions located elsewhere might actually harbor greater potential for coadaptation of genes. An obvious next step toward identification of key "speciation genes" is to identify the full range, location, and size of islands present in locations at the western edge of *A. gambiae*'s range, such as Guinea-Bissau, where interform gene flow is most substantial. A significant advantage of our study was that the relatively moderate cost of our arrays permitted large sample sizes to be genotyped individually. This provided sufficient statistical power to detect divergence at a single base resolution, which enabled detection of potentially small genomic islands that would likely be missed in a lower resolution analysis. Moreover, ascertainment bias toward either molecular form should have been limited since all SNPs on the array were known to be polymorphic within both M and S forms. However, the major counterbalancing disadvantage is that our genotyping array provided relatively low density coverage and included some substantial gaps such as that noted previously toward the centromere of chromosome 3L. Given the uncertain size and location of speciation islands and the generally very limited linkage disequilibrium observed in the *A. gambiae* genome (Harris et al. 2010; Weetman et al. 2010), whole genome resequencing would seem a clear best option but with the important caveat that study power should be sufficient to permit reasonably fine-scaled resolution. Our results (and those of Turner et al. 2005, described above) suggest that low background recombination alone is insufficient to explain the divergence of genomic islands, but the relative roles of reduced recombination and selection remain unclear, both in *A. gambiae* and in speciation genomics more generally (Noor and Bennett 2009; Nosil et al. 2009b). Therefore, an important goal will be to produce improved recombination maps for the *A. gambiae* genome, which at present are based on low resolution microsatellite data (Zheng et al. 1996; Pombi et al. 2006) and contain large regions for which no information is available.

The level of reproductive isolation between molecular forms in Guinea-Bissau is dramatically different to that observed in more easterly locations. Caputo et al. (2011) suggested that incompletely reproductively isolated forms might have come into secondary contact recently in Guinea-Bissau, perhaps as a result of recent invasion of the area by more ecologically generalist S forms. Our data appear entirely consistent with this hypothesis. Integrity of putatively adaptive genomic regions in the M forms persists to a large extent despite high and bidirectional gene flow, and there is a much greater disparity in sympatric versus allopatric differentiation between M and S than observed in Cameroon and Ghana (particularly evident at the 2L genomic island; see fig. 5). Nevertheless, despite considerable resilience of the genomic islands to gene flow in M forms, the broad scale homogeneity of the molecular forms elsewhere in the genome suggests that reproductive isolation may be in the process of breaking down. More direct evidence that this might be occurring is very limited at present. Caputo et al. (2011) found that linkage disequilibrium between diagnostic markers for the 3L and X chromosome islands was weaker in a 2007 collection from Guinea-Bissau than in those, which like ours, were from the mid-1990s; consistent with breakdown of reproductive barriers over time. How the rate and direction of genomic change may impact upon disease transmission is unknown, and the answer awaits improvement of the phenotypic and ecological information on *A. gambiae* in Guinea-Bissau. Nevertheless, the level of gene flow in Guinea-Bissau appears sufficiently high that shifts in divergence of M and S genomes could be detectable within an ecological rather than evolutionary timeframe and are thus highly amenable for study.

Supplementary Material

Supplementary figures s1–s3 and tables s1–s3 are available at *Molecular Biology and Evolution* online (http:// www.mbe.oxfordjournals.org/)

Acknowledgments

We thank Pie Müller, Emma Warr, and Sara Mitchell (Vector Group, LSTM), Frédéric Simard (IRD, Montpellier), Mohammed Chouaibou and Phillipe Nwane (OCEAC, Yaoundé, Cameroon), Alexander Egyir-Yawson (BNARI, Accra, Ghana), Loyce Okedi, Khyssa, Edward, Moses, James, Steven, and Andrew (NALIRRI, Tororo, Uganda), and Katinka Pälsson and Thomas G.T. Jaenson (Uppsala University, Sweden) for help with field collections. The manuscript was improved by comments from three anonymous reviewers on an earlier version. This work was funded primarily by the Innovative Vector Control Consortium, with additional support from National Institutes of Health grant R01AI082734-01.

References

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. 2008. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics* 9:323.

Ayala FJ, Coluzzi M. 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. *Proc Natl Acad Sci U S A.* 102(Suppl 1):6535–6542.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263–265.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond B.* 263: 1619–1626.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 57:289–300.

Butlin RK. 2010. Population genomics and speciation. *Genetica* 138:409–418.

Caputo B, Nwakanma D, Jawara M, Adiamoh M, Dia I, Konate L, Petrarca V, Conway DJ. 2008. Anopheles gambiae complex along The Gambia river, with particular reference to the molecular forms of An. gambiae s.s. *Malar J.* 7:182.

Caputo B, Santolamazza F, Vicente JL, et al. (14 co-authors). 2011. The "far-west" of Anopheles gambiae molecular forms. *PLoS One* 6:e16415.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. 2002. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* 298:1415–1418.

Corander J, Marttinen P. 2006. Bayesian identification of admixture events using multi-locus molecular markers. *Mol Ecol.* 15: 2833–2843.

Corander J, Marttinen P, Mäntyniemi S. 2006. Bayesian identification of stock mixtures from molecular marker data. *Fish Bull.* 104: 550–558.

Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 9:539.

Costantini C, Ayala D, Guelbeogo WM, et al. (12 co-authors). 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. *BMC Ecol.* 9:16.

Dabiré KR, Diabaté A, Namountougou M, Toé KH, Ouari A, Kengne P, Bass C, Baldet T. 2009. Distribution of pyrethroid and DDT resistance and the L1014F kdr mutation in *Anopheles gambiae s.l.* from Burkina Faso West Africa. *Trans R Soc Trop Med Hyg.* 103:1113–1120.

della Torre A, Fanello C, Akobeto M, Dossou-yovo J, Favia G, Petrarca V, Coluzzi M. 2001. Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa. *Insect Mol Biol.* 10:9–18.

della Torre A, Tu Z, Petrarca V. 2005. On the distribution and genetic differentiation of *Anopheles gambiae s.s.* molecular forms. *Insect Biochem Mol Biol.* 35:755–769.

Diabate A, Dabire RK, Millogo N, Lehmann T. 2007. Evaluating the effect of postmating isolation between molecular forms of *Anopheles gambiae* (Diptera: Culicidae). *J Med Entomol.* 44:60–64.

Diabate A, Dao A, Yaro AS, Adamou A, Gonzalez R, Manoukis NC, Traoré SF, Gwadz RW, Lehmann T. 2009. Spatial swarm segregation and reproductive isolation between the molecular forms of Anopheles gambiae. *Proc Biol Sci.* 276:4215–4222.

Djogbénou L, Chandre F, Berthomieu A, Dabiré R, Koffi A, Alout H, Weill M. 2008. Evidence of introgression of the *ace-1*^{R}. mutation and of the *ace-1* duplication in West African *Anopheles gambiae s.s.* *PLoS One* 3:e2172.

Esnault C, Boulesteix M, Duchemin JB, et al. (12 co-authors). 2008. High genetic differentiation between the M and S molecular forms of *Anopheles gambiae* in Africa. *PLoS One* 3:e1968.

Etang J, Fondjo E, Chandre F, Morlais I, Brengues C, Nwane P, Chouaibou M, Ndjemai H, Simard F. 2006. First report of knockdown mutations in the malaria vector *Anopheles gambiae* from Cameroon. *Am J Trop Med Hyg.* 74:795–797.

Etang J, Vicente JL, Nwane P, Chouaibou M, Morlais I, Do Rosario VE, Simard F, Awono-Ambene P, Toto JC, Pinto J. 2009. Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae s.s.* populations from Cameroon with emphasis on insecticide knockdown resistance mutations. *Mol Ecol.* 18:3076–3086.

Fanello C, Santolamazza F, della Torre A. 2002. Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP. *Med Vet Entomol.* 16:461–464.

Felsenstein J. 2008. PHYLIP. Phylogeny inference package. version 3.68. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Gentile G, Slotman M, Ketmaier V, Powell JR, Caccone A. 2001. Attempts to molecularly distinguish cryptic taxa in Anopheles gambiae s.s. *Insect Mol Biol.* 10:25–32.

Gimmoneau G, Bouyer J, Morand S, Besansky NJ, Diabate A. 2010. A behavioral mechanism underlying ecological divergence in the malaria mosquito. Anopheles gambiae. *Behav Ecol.* 21:1087–1092.

Hänfling B, Weetman D. 2006. Concordant genetic estimators of migration reveal anthropogenically enhanced source-sink population structure in the river sculpin, Cottus gobio. *Genetics* 173:1487–1501.

Harris C, Rousset F, Morlais I, Fontenille D, Cohuet A. 2010. Low linkage disequilibrium in wild *Anopheles gambiae* s.l. populations. *BMC Genet.* 11:81.

Holt RA, Subramanian GM, Halpern A, et al. 123 co-authors. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* 298:129–149.

Lanzaro GC, Touré YT, Carnahan J, Zheng L, Dolo G, Traoré S, Petrarca V, Vernick KD, Taylor CE. 1998. Complexities in the genetic structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis. *Proc Natl Acad Sci U S A.* 95:14260–14265.

Lawniczak MKN, Emrich S, Holloway AK, et al. (30 co-authors). 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330:512–514.

Lehmann T, Hawley WA, Grebert H, Collins FH. 1998. The effective population size of *Anopheles gambiae* in Kenya: implications for population structure. *Mol Biol Evol.* 15:263–276.

Lehmann T, Diabate A. 2008. The molecular forms of *Anopheles gambiae*: a phenotypic perspective. *Infect Genet Evol.* 8:737–746.

Lehmann T, Licht M, Elissa N, Maega BT, Chimumbwa JM, Watsenga FT, Wondji CS, Simard F, Hawley WA. 2003. Population Structure of *Anopheles gambiae* in Africa. *J Hered.* 94:133–147.

Lynd A, Weetman D, Barbosa S, Egyir-Yawson A, Mitchell S, Pinto J, Hastings I, Donnelly MJ. 2010. Field, genetic and modelling approaches show strong positive selection acting upon an insecticide resistance mutation in Anopheles gambiae s.s. *Mol Biol Evol.* 27:1117–1125.

McBride CS, Singer MC. 2010. Field studies reveal strong postmating isolation between ecologically divergent butterfly populations. *PLoS Biol.* 8:e1000529.

Miller M. 1997. RxC a Windows program for the analysis of contingency tables via the metropolis algorithm. Accessed August 15, 2010 from: http://www.marksgeneticsoftware.net/rxc.htm

Müller P, Warr E, Stevenson BJ, et al. (13 co-authors). 2008. Field-caught permethrin-resistant *Anopheles gambiae* overexpress CYP6P3, a P450 that metabolises pyrethroids. *PLoS Genet.* 4:e1000286.

Ndiath MO, Brengues C, Konate L, Sokhna C, Boudin C, Trape JF, Fontenille D. 2008. Dynamics of transmission of *Plasmodium falciparum* by *Anopheles arabiensis* and the molecular forms M and S of *Anopheles gambiae* in Dielmo, Senegal. *Malar J.* 7:136.

Neafsey DE, Lawniczak MK, Park DJ, et al. (17 co-authors). 2010. SNP genotyping defines complex gene flow boundaries among sympatric African malaria vector mosquitoes. *Science* 330:514–517.

Noor MA, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103:439–444.

Norris DE, Shurtleff AC, Touré YT, Lanzaro GC. 2001. Microsatellite DNA polymorphism and heterozygosity among field and laboratory populations of Anopheles gambiae ss (Diptera: Culicidae). *J Med Entomol.* 38:336–340.

Nosil P, Funk DJ, Ortiz-Barrientos D. 2009a. Divergent selection and heterogeneous genomic divergence. *Mol Ecol.* 18:375–402.

Nosil P, Harmon LJ, Seehausen O. 2009b. Ecological explanations for incomplete speciation. *Trends Ecol Evol.* 24:145–156.

Oliveira E, Salgueiro P, Palsson K, Vicente JL, Arez AP, Jaenson TG, Caccone A, Pinto J. 2008. High levels of hybridization between molecular forms of *Anopheles gambiae* from Guinea Bissau. *J Med Entomol.* 45:1057–1063.

Pennetier C, Warren B, Dabire KR, Russell IJ, Gibson G. 2010. "Singing on the wing" as a mechanism for species recognition in the malarial mosquito Anopheles gambiae. *Curr Biol.* 20:131–136.

Pombi M, Caputo B, Simard F, Di Deco MA, Coluzzi M, della Torre A, Costantini C, Besansky NJ, Petrarca V. 2008. Chromosomal plasticity and evolutionary potential in the malaria vector *Anopheles gambiae* sensu stricto: insights from three decades of rare paracentric inversions. *BMC Evol Biol.* 8:309.

Pombi M, Stump AD, Della Torre A, Besansky NJ. 2006. Variation in recombination rate across the X chromosome of Anopheles gambiae. *Am J Trop Med Hyg.* 75:901–903.

Rambault A. 2006. Fig. tree. Accessed June 12, 2010 from: http://tree.bio.ed.ac.uk/software/figtree

Rousset F. 2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol Ecol Resources* 8:103–106.

Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, della Torre A. 2008a. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. *Malar J.* 7:163.

Santolamazza F, Calzetta M, Etang J, Barrese E, Dia I, Caccone A, Donnelly MJ, Petrarca V, Simard F, Pinto J, della Torre A. 2008b. Distribution of knock-down resistance mutations in *Anopheles gambiae* molecular forms in west and west-central Africa. *Malar J.* 7:74.

Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, Smith CD, Sharakhov IV. 2010. Genome mapping and characterization of the *Anopheles gambiae* heterochromatin. *BMC Genomics* 11:459.

Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, Fotsing JM, Fontenille D, Besansky NJ, Costantini C. 2009. Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol.* 9:17.

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science* 236:787–792.

Slotman MA, Reimer LJ, Thiemann T, Dolo G, Fondjo E, Lanzaro GC. 2006. Reduced recombination rate and genetic differentiation between the M and S forms of Anopheles gambiae s.s. *Genetics* 174:2081–2093.

Stump AD, Fitzpatrick MC, Lobo NF, Traoré S, Sagnon N, Costantini C, Collins FH, Besansky NJ. 2005. Centromere-proximal differentiation and speciation in Anopheles gambiae. *Proc Natl Acad Sci U S A.* 102:15930–15935.

Taylor C, Touré YT, Carnahan J, Norris DE, Dolo G, Traoré SF, Edillo FE, Lanzaro GC. 2001. Gene flow among populations of the malaria vector, *Anopheles gambiae*, in Mali, West Africa. *Genetics* 157:743–750.

Tripet F, Touré YT, Taylor CE, Norris DE, Dolo G, Lanzaro GC. 2001. DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of Anopheles gambiae. *Mol Ecol.* 10:1725–1732.

Turner TL, Hahn MW. 2007. Locus- and population-specific selection and differentiation between incipient species of Anopheles gambiae. *Mol Biol Evol.* 24:2132–2138.

Turner TL, Hahn MW. 2010. Genomic islands of speciation or genomic islands and speciation? *Mol Ecol.* 19:848–850.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3:e285.

Via S. 2009. Natural selection in action during speciation. *Proc Natl Acad Sci U S A.* 106 (Suppl 1):9939–9946.

Via S, West J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol Ecol.* 17:4334–4345.

Weetman D, Wilding CS, Steen K, Morgan JC, Simard F, Donnelly MJ. 2010. Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilibrium genome. *PLoS One* 5:e13140.

Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, Pasteur N, Guillet P, Raymond M. 2000. The kdr mutation occurs in the Mopti form of *Anopheles gambiae s.s.* through introgression. *Insect Mol Biol.* 9:451–455.

White BJ, Cheng C, Sangaré D, Lobo NF, Collins FH, Besansky NJ. 2009. The population genomics of trans-specific inversion polymorphisms in Anopheles gambiae. *Genetics* 183:275–288.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ. 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of. Anopheles gambiae. *Mol Ecol.* 19:925–939.

Wilding CS, Weetman D, Steen K, Donnelly MJ. 2009. High, clustered, nucleotide diversity in the genome of *Anopheles gambiae* revealed by SNP discovery through pooled-template sequencing: implications for high-throughput genotyping protocols. *BMC Genomics* 10:320.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Wu C-I. 2001. The genic view of the process of speciation. *J Evol Biol.* 14:851–865.

Wu CI, Ting CT. 2004. Genes and speciation. *Nat Rev Genet.* 5:114–122.

Yawson AE, McCall PJ, Wilson MD, Donnelly MJ. 2004. Species abundance and insecticide resistance of *Anopheles gambiae* in selected areas of Ghana and Burkina Faso. *Med Vet Entomol.* 18:372–377.

Yawson AE, Weetman D, Wilson MD, Donnelly MJ. 2007. Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector *Anopheles gambiae s.s.* in Ghana. *Genetics* 175:751–761.

Zheng L, Benedict MQ, Cornell AJ, Collins FH, Kafatos FC. 1996. An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae. *Genetics* 143:941–952.